



# Correlation within SNCF administrative regions among track segment maintenance cost equation residuals of a country-wide model

Marc Gaudry, Emile Quinet

## ► To cite this version:

Marc Gaudry, Emile Quinet. Correlation within SNCF administrative regions among track segment maintenance cost equation residuals of a country-wide model. 2015. halshs-01112249

**HAL Id: halshs-01112249**

**<https://shs.hal.science/halshs-01112249>**

Preprint submitted on 2 Feb 2015

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



**PARIS SCHOOL OF ECONOMICS**  
ÉCOLE D'ÉCONOMIE DE PARIS

**WORKING PAPER N° 2015 – 04**

**Correlation within SNCF administrative regions among track  
segment maintenance cost equation residuals of a country-wide  
model**

**Marc Gaudry  
Émile Quinet**

**JEL Codes: C61, D61, D92, H54, R41, R48**

**Keywords: rail infrastructure, current maintenance cost, rail track degradation,  
Box-Cox transformation, directed autocorrelation, spatial autocorrelation,  
SNCF, French rail network, network data by segment, SNCF administrative  
regions, regional differences**



**PARIS-JOURDAN SCIENCES ÉCONOMIQUES**

48, Bd JOURDAN – E.N.S. – 75014 PARIS  
TÉL. : 33(0) 1 43 13 63 00 – FAX : 33 (0) 1 43 13 63 10  
[www.pse.ens.fr](http://www.pse.ens.fr)

# Correlation within SNCF administrative regions among track segment maintenance cost equation residuals of a country-wide model

Marc Gaudry<sup>1</sup> and Émile Quinet<sup>2</sup>



1 Agora Jules Dupuit (AJD)  
Université de Montréal  
Montréal, [marc.gaudry@umontreal.ca](mailto:marc.gaudry@umontreal.ca)



2 Paris-Jourdan Sciences Économiques (PSE)  
École des Ponts ParisTech (ENPC)  
Paris, [quinet@enpc.fr](mailto:quinet@enpc.fr)

The authors thank Cong-Liem Tran for computing assistance, Gérard Mathieu for comments, and are grateful to *Société nationale des chemins de fer français* (SNCF) for financial support and for allowing use of a database constructed by Pascaline Boyer. Forthcoming in Funck, R., Rothengatter, W. (Hrsg.), *Man, Environment, Space and Time - Economic Interactions in Four Dimensions. Karlsruhe Papers in Economic Policy Research*, Volume 34, Nomos Verlag, p. 207-232, 2014.

JEL classification codes: C61, D61, D92, H54, R41, R48

**Université de Montréal**  
Agora Jules Dupuit - Publication AJD-145

-----  
**Paris-Jourdan Sciences Économiques**  
PSE Working Paper N° -2015-04

Version 4, November 7, 2013; revised 9<sup>th</sup> April, 2014.

## **Abstract**

Using as reference a recent France-wide model of rail infrastructure maintenance cost where regression model residuals associated to track segments are assumed to be similarly correlated among themselves within the 23 administrative regions of the national firm (SNCF), we attempt to explain the presence of the strong positive and stationary correlation coefficient estimated in this manner and to probe the extent to which the assumption of a common correlation coefficient across administrative regions might be refined and interpreted.

We first find that country-wide within-region correlation among residuals is not weakened if regional dummy variables are added to the model in the hope of finding interpretative clues, notably of the presence of climate effects unaccounted for in the model specification. The implicit region-specific weather effects indirectly so represented turn out to be extremely weak, if present at all, and do not affect the strength of extant within-region correlation or the need to make sense of it.

We then explore differences in correlation coefficient estimates among regions and show that, within the reference maintenance cost model, two large geographic groupings of regions, each comprising in the East or in the North about 15% of total available track segments, in fact have own residuals that are uncorrelated among themselves, in contrast to the more numerous 70% of segment residuals remaining in the rest of regions, which as a group remain robustly and positively correlated, and in a stationary manner. As the two groupings with uncorrelated residuals closely match the networks of regional firms merged into the SNCF conglomerate in 1938, we hypothesize, *faute de mieux* and in the absence for the moment of refined local climate variables to pursue unpromising missing variable tests, that within-firm accounting traditions might have survived centralized management control of the assignment of track surveillance, maintenance and repair invoices to track sections centrally defined for accounting management purposes.

**Keywords:** rail infrastructure, current maintenance cost, rail track degradation, Box-Cox transformation, directed autocorrelation, spatial autocorrelation, SNCF, French rail network, network data by segment, SNCF administrative regions, regional differences.

## **Table of contents**

1. The context: European rail infrastructure maintenance cost modeling.....	3
2. Background of the model selected for tests of spatial correlation .....	3
2.1. First wave rail track segment models.....	3
2.2. A particular second wave track segment model.....	5
3. Basic data and results of the reference maintenance cost model .....	7
4. Within-region correlation among regression errors .....	10
5. Conclusion .....	14
6. References .....	14

## **List of tables**

Table 1. Mean values of principal variables used in reference model.....	7
Table 2. BCT forms and elasticities of maintenance cost $m$ with respect to 8 regressors .....	8
Table 3. SNCF regions: neighbors (nb), segments (sg) in the sample and $t$ -statistic of dummies .....	9
Table 4. Four geographic partitions of matrix $R_1$ into pairs of complementary matrices.....	11
Table 5. Testing geographic and sociological decompositions of matrix $R_1$ .....	12
Table 6. Family structure in France and the Liberal/Authoritarian partition of SNCF regions.....	13

**Journal of Economic Literature: C61, D61, D92, H54, R41, R48**

## 1. The context: European rail infrastructure maintenance cost modeling<sup>1</sup>

Models of infrastructure maintenance costs estimated from rail track segment data, linking such costs to traffic and technical characteristics of ways, are not to be confused with models of total rail cost (including train service provision) estimated from total expenditures of railway firms.

The latter arose in the 1980's and 1990's in a context where the Trans-Log form (Christensen *et al.*, 1971a, 1971b) was progressively adopted as a standard tool of analysis of transport firm costs across all modes: representative rail examples include Caves *et al.* (1980a, 1980b, 1985) in the USA and Borger (1991, 1992) in Belgium. The former, sometimes separately explaining track renewal costs by track segment in addition to maintenance costs, arose specifically in Europe in the 2000's: early studies include Idström (2002), Johansson & Nilsson (2002, 2004) and Gaudry & Quinet (2003). Such studies are still few, European, and very rare for other modes: Ben-Akiva & Gopinath (1995), who studied road maintenance costs with Brazilian highway section data, provide a notable exception. We found no example of applications to airport runways or taxiways.

The model of the French rail network maintenance costs for the year 2007 used here to study spatial correlation of its residuals actually belongs to a second wave of rail maintenance cost models. First wave models do not impose complex behavioral conditions on the estimated cost function other than short-run cost minimization. Second wave models are more ambitious in this respect and assume on the part of the infrastructure firm some joint optimization of maintenance and regeneration expenditures: these stronger assumptions have implications for both the contents and the shape of the maintenance cost function.

It is on a function of this latter type that we will perform spatial and more general tests of “directed” correlation pointing both to unexplained similarities within administrative regions and to dissimilarities across such regions of the nation-wide SNCF firm. As the interpretation of extant region-specific correlation among residuals is no less uncertain than that of serial correlation, our analysis will only speculate on the causes of uncovered correlations.

## 2. Background of the model selected for tests of spatial correlation

### 2.1. First wave rail track segment models

In the first wave of studies based on rail track segment data, a behavioral assumption of short run cost minimization by infrastructure owners is typically made and the cost function is estimated with an eye for its optimal pricing implications, especially in countries where the marginal cost pricing doctrine of the European Union Commission is applied. A representative function might be:

$$(1-A) \quad m_t = f(K(h_0, e_0), q(n, w), S(h))_t,$$

where, for a given year  $t$ , the maintenance cost per unit of track segment  $m$  is a function of its technical characteristics  $K$ , current traffic  $q$  and latent deterioration indicators  $S(h)$ . A particularly complete set of recent models within this stream is found in the numerous country reports of the Cost Allocation of Transport INfrastructure cost (CATRIN) Consortium, summarized in Wheat *et al.* (2009), and in related contributions (*e.g.* Andersson *et al.*, 2012).

Note that Eq. (1-A) contains many sub-sets of explanatory variables found in these various country studies:  $K$  is split between planned reference quality or “standing”  $h_0$  (such as maximum allowed

---

<sup>1</sup> The authors thank Cong-Liem Tran for computing assistance, Gérard Mathieu for comments, and are grateful to *Société nationale des chemins de fer français* (SNCF) for financial support and for allowing use of a database constructed by Pascaline Boyer.

design speed  $v$ ) and state characteristics  $e_0$  (such as segment length; circuitousness; number of tracks, switches and tunnels; electrification; centralized automated traffic control; *etc.*); also, current traffic  $q$  is split between the number of currently circulating trains ( $n$ ) and their weight ( $w$ ); finally, current deterioration performance indicators  $S(h)$  include many possible other factors (such as age of rails and of sleepers; proportion of wood sleepers; *etc.*). We have found no rail paper using an actual measure of track performance or condition like the Longitudinal Deformation index appearing in the model of interest below: the  $h$  in question in (1-A) are only the latent determinants of unobserved track condition  $S$ , not direct measures of it as found below.

As the index  $t$  in (1-A) refers in practice to very few consecutive years of available data, typically one and at most three, the brevity of large track segment panels makes it possible, if not necessary, for the above early and for later authors to abstract entirely from input prices: they are then assumed constant in space (and during the very short time periods available) for the given national infrastructure provider studied, for instance Swedish in Andersson (2006), Swiss in Marti & Neuenschwander (2006) and Austrian in Link (2009) — all three of which use CES (Log-Log) specifications for  $f(\cdot)$ , as most early authors did.

On this last point of functional form, CATRIN mandated simple Box-Cox (SBC) tests, consisting in using a single Box-Cox transformation (BCT), defined in (2-D) below, on the dependent and independent variables of a CES reference model: all resulting national model results so obtained (*e.g.* Andersson, 2009; Link, 2009; Marti *et al.*, 2009) proved far superior to the references ones.

But the most extensive Box-Cox tests carried out within CATRIN showed that, although the popular Trans-Log function was not superior to the CES function for  $f(\cdot)$ , CES-inspired generalizations named Unrestricted-Generalized Box-Cox (U-GBC) — and even the single-BCT Restricted (R-GBC) specification inspired by Berndt & Khaled (1979)<sup>2</sup> — were, statistically speaking, “infinitely” superior to both CES and Trans-Log forms. These new forms (Gaudry & Quinet, 2009a, 2009b), already tried in the 2003 round of tests by the same authors, were:

$$(2-A) \quad \begin{aligned} y^{(\lambda)} &= \beta_0 + \sum_{k=1}^{k=r} \beta_k X_k^{(\lambda)} + \sum_{i=1}^{i=r} \sum_{j=1}^{j=r} \beta_{ij} (X_i X_j)^{(\lambda)} , & i \neq j & \quad \text{R-GBC} \\ y^{(\lambda_y)} &= \beta_0 + \sum_{k=1}^{k=r} \beta_k X_k^{(\lambda_k)} + \sum_{i=1}^{i=r} \sum_{j=1}^{j=r} \beta_{ij} (X_i X_j)^{(\lambda_{ij})} , & i \neq j & \quad \text{U-GBC} \end{aligned}$$

where  $i \neq j$  because BCT are invariant to a power transformation (Gaudry & Laferrière, 1989) and, in consequence, the BCT of a squared variable is not distinct from that of the simply transformed variable. As many as 13 BCT were used in various model runs, including some with five  $\lambda_k$  specific to train types demonstrating the existence of differences in track damages across train types<sup>3</sup>.

Remembering that prices are absent from these cross-sectional models, the U-GBC and R-GBC tests in fact showed that the Trans-Log is a very poor model of physical interaction between track and traffic because this interaction is demonstrably not Log-Log. They also showed, by systematically studying the impact of total traffic disaggregation among train types that, if certain precautions were taken<sup>4</sup>, the BCT was the best tool to test the forms of variables that include some zero observations (*e.g.* not all types of trains use all track segments), and was in fact superior, in theory and practice, to both CES (Log-Log) and Trans-Log specifications in handling zero values.

<sup>2</sup> They applied a single common BCT to all individual variables of a Trans-Log specification.

<sup>3</sup> As these models contain no speed variable by train type, results by train type combines weight, speed and other effects of actual weights per axle, un-sprung mass and technical features absorbing the effect of gross axle weights.

<sup>4</sup> Unfortunately, many previous applications of BCT to zero observations, such as Caves *et al.* (1980a, 1985), do not take these precautions and, as a result, estimate values of BCT dependent on the units of measurement of the (traffic) variables containing the zeroes. This fault makes their BCT form estimates uninteresting and unusable for readers.

## 2.2. A particular second wave track segment model

**Economic specification of a three-phase model.** If the assumption of short-run cost minimization is replaced by that of joint optimization of maintenance and renewal, the cost function is enriched by new terms: in particular, the cumulative value of traffic  $Q$  (in terms of number of trains  $N$  or their weight  $W$ )<sup>5</sup> and  $t$ , time elapsed since the last regeneration, now matter in principle.

The specifics of the particular source model considered for our tests (Gaudry *et al.*, 2013), from which we will borrow heavily in this derived paper, also include some dynamic control variables  $c$ , functions of the firm's track quality target  $Sc$ , because deviations of actual track conditions  $S$  from their target levels bring forth corrections of maintenance expenditures. Noting that the latent condition indicators  $S(h)$  of (1-A) are replaced by actual condition  $S$  and target condition  $Sc$ , the doubly enriched cost function may now be written:

$$(1-B) \quad m_t = f(K(h_0, e_0), q(n, w), Q(N, W), t, c(S, Sc))_t.$$

Indeed, the actual theoretical model at hand in fact defines three maintenance phases. In the first and third phases, *i.e.* after a segment has been renewed and before it is to be renewed, maintenance is *nil* or very low and limited to surveillance; but during the central cruising phase, maintenance is performed in such a way as to follow the target track quality or condition path. This means that, for the estimation of determinants of maintenance, only this central phase is relevant: one then has to remove from the available sample observations belonging to unrepresentative first and third phases.

**Econometric specification with and without a  $\cap$ -shaped expenditure time profile.** This is effected in a first step by estimating with the full sample, within a given multivariate U-GBC specification, the inverse U-shaped<sup>6</sup> time profile of maintenance expenditures associated with one particular regressor, cumulative traffic  $Q$  (here the cumulative weight of trains  $W$  since the last track segment regeneration):

$$(2-A) \quad Q(W_t) = \beta_{Q1} W_t^{(\lambda_{Q1})} + \beta_{Q2} W_t^{(\lambda_{Q2})}.$$

In the second step, very small and very large values of observations of this variable are progressively removed until the cost function is monotonic in terms of  $W$  because the two BCT powers assigned to the same variable  $W$  converge to a unique value and (2-A) becomes monotonic.

With the doubly censored sample, we then tested a number of U-GBC specifications containing interactions between *cumulative traffic* and some measure of *maximum allowed speed*, for both total traffic and traffic by train type, but none of the interactions proved significant under the new specification enriched by cumulative traffic ( $W$ ) and control variables ( $S, S^*$ ). In the absence of interactions and of non monotonic regressors, these U-GBC specifications reduce to a standard multivariate BCT generalization of the CES which may now be written for our spatial correlation tests with an explicit residual error  $v_t$  associated with each of the  $T$  segments in the sample:

$$(2-B) \quad y_t^{(\lambda_y)} = \beta_0 + \sum_k \beta_k X_{kt}^{(\lambda_k)} + v_t$$

$$(2-C) \quad v_t = \sum_{\ell=1}^{\ell=2} \rho_{\ell} \left( \sum_{n=1}^{n=T} \tilde{r}_{\ell,n} v_n \right) + w_t$$

where, in the first equation duly containing an intercept (Schlesselman, 1971) to guarantee the invariance of form parameters to changes in the units of measurement of the  $X_k$ , the Box-Cox

<sup>5</sup> As did a cumulative traffic variable used in the *ad hoc* road example mentioned above.

<sup>6</sup> For details on the use of two BCT on the same variable of a regression, see Gaudry *et al.* (2000) who show that the procedure makes it possible to model asymmetric turning forms as well as classical symmetric quadratic U shapes.

transformation (BCT) applicable to any strictly positive<sup>7</sup> variable  $Var_v$  is commonly defined without Tukey's shift parameter as:

$$(2-D) \quad Var_v^{(\lambda_v)} \equiv \begin{cases} [(Var_v)^{\lambda_v} - 1]/\lambda_v, & \lambda_v \neq 0, \\ \ln(Var_v) & , \quad \lambda_v \rightarrow 0, \end{cases}$$

and where, in the second equation,  $\tilde{r}_{\ell,m}$  denotes the typical element of matrix  $\tilde{R}_\ell$ , a row or column-normalized square ( $T$  by  $T$ ) Boolean matrix  $R_\ell$  expressing hypothesis  $\ell$  concerning the presence of correlation among designated residuals, which may then behave as substitutes ( $\rho_\ell > 0$ ) or complements ( $\rho_\ell < 0$ ) to the residual of segment  $t$ . And rewriting (2-C) in matrix notation eases the explicitation of matrix  $\tilde{R}_\ell$ , namely

$$(2-E) \quad v = \sum_{\ell=1}^{\ell=2} \rho_\ell \tilde{R}_\ell v + w, \text{ with } \tilde{R}_\ell = \pi_\ell [I - (1 - \pi_\ell) \bar{R}_\ell]^{-1} \bar{R}_\ell, \quad (0 < \pi_\ell \leq 1),$$

where the new proximity parameter  $\pi_\ell$  measures the relative influence of “near” and “distant” neighbors of  $R_\ell$ , as soon discussed below.

**Directed autocorrelation.** Matrix  $\tilde{R}_\ell$  in (2-E) results from three steps. In the first, a square matrix  $R_\ell$  is defined with typical element  $r_{\ell,m}=1$ , to express an hypothesized correlation between any two residuals  $v_t$  and  $v_n$  (and  $r_{\ell,m}=0$  otherwise). It is often called a contiguity matrix because, historically, spatial hypotheses were first used to define it; but clearly any criterion — spatial, temporal or socio-economic — may specify a pattern of interdependence among residuals: the researcher's *directed* choice<sup>8</sup> is not limited by the natural order of the data.

In the second step, it has long been the practice (Ord, 1975) to row or column normalize this matrix and Bolduc (1987) has shown that the resulting normalized matrix  $\bar{R}_\ell$  guarantees a convex likelihood function over the stable unit interval of  $\rho_\ell$ . This Boolean formulation is « idiot-proof » in the sense that, based only on zeroes and ones, it avoids the complications of lack of invariance of estimates to changes in units of measurement that arise when the elements of  $R_\ell$  are functions of continuous variables, such as distance or income (Bolduc *et al.*, 1989, p. 369). But something might then be needed to compensate for the discreteness of  $R_\ell$ , as many distributed phenomena are likely to be smooth and their representation by a few « all-or-nothing » Boolean slices insufficient: the solution resides in taking due account of the whole set of near and distant “neighbor” slices, as long done with distributed lags in time-series analysis.

**Contiguous and distributed “spatial” lags.** This is precisely the idea of the Blum *et al.* (1995/1996) approach where powers of  $\bar{R}_\ell$  generate a sequence of contiguity matrices which define degrees of neighborliness or proximity ( $\bar{R}_\ell^2, \dots, \bar{R}_\ell^c, \dots$ ), with  $\bar{R}_\ell^2$  denoting neighbors of neighbors, and  $\bar{R}_\ell^c$  higher powers.

Under the assumption that the impact of these close and distant neighbors decreases geometrically with «distance»  $c$ , as in Koyck (1954) distributed lags of time-series (whence the name

<sup>7</sup> The dummy variable and 0-replacement methods available to transform variables that contain some 0 are discussed and used in Gaudry & Quinet (2010). Here 0 traffic values are replaced by 0,00001, unless stated otherwise — for instance in the detection of Phase C segments where the dummy variable method is used.

<sup>8</sup> For an example of directed choice involving correlation of residuals across socio-economic groups, see Gaudry & Blum (1988). For distributed processes in transport and trade flow models, see Gaudry (2004).



*Autoregressive Contiguous Distributed* (AR-C-D) for this analogue process), one obtains  $\tilde{R}_\ell = \pi_\ell [I - (1 - \pi_\ell)\bar{R}_\ell]^{-1} \bar{R}_\ell$ , where the new proximity parameter  $\pi_\ell$  allows for endogenization of the relative importance of near and distant effects. If  $\pi_\ell = 1$ ,  $\tilde{R}_\ell = \bar{R}_\ell$ , in which case only the adjacent neighbors have an impact on the correlations among the residuals selected by the “residue impact criterion” expressed in  $R_\ell$ , exactly as in the classical (Ord, 1975; Cliff & Ord, 1981) case. By contrast, as  $\pi_\ell \rightarrow 0$ , the near effect is reduced to a minimum in favor of the distant effect. In this R-Koyck formulation, the parameter  $\pi_\ell$  therefore generally weighs the *relative* importance of near and distant effects, *i.e.* the sharpness or slope of the decline.

Using the L-2.1 algorithm (Tran & Gaudry, 2008) to jointly estimate parameters of Eq. (2-B)-(2-C), our analysis of maintenance cost by segment  $m_t$  tests a number of specifications of the  $R_\ell$  matrix and refers to further tests with two distributed processes  $\pi_\ell$  reported in the source paper.

### 3. Basic data and results of the reference maintenance cost model

Table 1 presents the main data of the uncensored sample of 700 observations, representing about 70% of total network segments, a number reduced to 580 segments in the cost estimation model.

**Table 1. Mean values of principal variables used in reference model**

Variables		(2007 sample size 700)			
m <sub>t</sub>	Current maintenance expenses				
	total cost per km (current Euros)		52 432		
	surveillance		14 821		
	maintenance		37 611		
e <sub>0</sub>	Technical state variables K				
	length of segment (meters)		28 219 <sup>1</sup>		
	length of all tracks (meters)		46 186		
	number of switches per segment		25,48		
h <sub>0</sub>	Initial standing				
	maximum allowed speed (km/h)		114,69		
t <sub>t</sub>	Current condition				
	average age of rails (years)		30,56		
S <sub>t</sub>	Service quality of the track				
	NL index of longitudinal track rectitude (mm)		1,451 <sup>2</sup>		
q <sub>t</sub>	Traffic per day: trains and gross tons		trains <sup>3</sup>	weight <sup>3</sup>	(per train)
	GL : long distance passenger trains (VFE)		17,99	10 255	(570)
	TGV: high speed trains		9,25	5 612	(601)
	Classic intercity trains (Corail)		8,74	4 643	(531)
	TER : regional passenger trains		24,44	5 559	(228)
	IdF : Île-de-France passenger trains		16,21	6 215	(384)
	Fret : freight trains		13,70	15 484	(1131)
	HLP : locomotives		2,84	275	(97)
	Total for the six categories of trains		75,16	37 788	(503)

<sup>1</sup> Out of a total network of about 30 000 km. <sup>2</sup> Available over the period 2000-2010 for a subset of 608 observations. <sup>3</sup> Available over the period 1995-2007 for all 700 observations.

<sup>1</sup>Out of a total network of about 30 000 km. <sup>2</sup>Available over the period 2000-2010 for a subset of 608 observations. <sup>3</sup>Available over the period 1995-2007 for all 700 observations.

Table 2 presents the main variants of the maintenance cost reference model<sup>9</sup> estimated from the censored (cruise phase) data. We make four principal comments:

- specification of variables and flexible BCT forms.* Column 1 presents a starting formulation with a fixed multiplicative Log-Log form as close in contents as possible to that of the UK Office of Rail Regulation (ORR) function, described in Ongaro & Iwnicki (2009). Under this form specification, all signs are as expected except for the time proxy variable (*agerail*) negative sign which progressively corrects itself

<sup>9</sup> The elasticities found in Table 2 differ slightly from those found in the final model due to slight changes later made to some values of a particular regressor. This modification has no effect on spatial correlation test results.

with the form and content enrichments of the next three columns; but this form-sensitive variable of expected positive sign is never really significant and never obtains a strong elasticity.

Concerning fit, massive Log Likelihood gains of 111 points are obtained in Column 2 by estimating only 4 BCT, all in fact very different from the logarithmic starting values assumed in Column 1: 0,25 for the dependent variable  $m$ , 0,53 for the three technical state variables  $e$ , 0,63 for total tonnage (the sum of cumulative and current values ( $W+w$ )) and 2,11 for  $V$ , a measure of maximum operation speed constructed from assumed maximum speeds by category of train<sup>10</sup>  $c$  weighted by their shares of total traffic [cumulative ( $W_c$ ), measured at the beginning of each year, and current ( $w_c$ )] measured in tons.

**Table 2. BCT forms and elasticities of maintenance cost  $m$  with respect to 8 regressors**

Elasticity $\eta(m)$ , $\lambda$		Column	1		2		3		4	
$(t\text{-statistic}=0)^*$ , $[t\text{-statistic}=1]$			$\eta(m)$	$\lambda$	$\eta(m)$	$\lambda$	$\eta(m)$	$\lambda$	$\eta(m)$	$\lambda$
<b>m</b>	Total maintenance cost per km (dependent variable) $(t=0)$ $[t=1]$	n.a.		<b>0</b>	n.a.	<b>0,25</b> (19.38) [-59.68]	n.a.	<b>0,24</b> (18.18) [-57.33]	n.a.	<b>0,22</b> (16.83) [-59.32]
<b><math>\beta_0</math></b>	Intercept $(t=0)$	n.a. (11.60)			n.a. (9.50)		n.a. (8.62)		n.a. (9.98)	
<b>S</b>	Target $\Delta$ service : 2006-2005 [E(S <sub>ct-1</sub> )- E(S <sub>ct-2</sub> )] $(t=0)$						0.08** (4.51)	<b>1</b>	0.10** (5.22)	<b>1</b>
	Trajectory correction: (obs.-target) <sub>2005</sub> [S <sub>t-2</sub> - E(S <sub>ct-2</sub> )] $(t=0)$						-0.002 (-3.14)		-0.002 (-3.50)	
<b>e<sub>0</sub></b>	Segment length $(t=0)$	-0.121 (-2.91)	<b>1</b>	-0.097 (-2.35)	<b>0.53</b> (8.53) [-7.71]	-0.082 (-2.10)	<b>0.58</b> (8.81) [-6.29]	-0.038 (-0.98)	<b>0.50</b> (8.39) [-8.48]	
	Track length $(t=0)$	0.116 (1.05)		0.178 (2.32)		0.149 (1.93)		0.181 (2.37)		
	Number of switches $(t=0)$	0.191 (10.01)		0.416 (20.21)		0.394 (19.53)		0.440 (19.94)		
<b>(W+w)</b>	Cumulative+ current total tons (W+w) $(t=0)$ $[t=1]$	0.279 (8.21)	<b>0</b>	0.251 (9.35)	<b>0.63</b> (2.93) [-1.72]	0.273 (10.39)	<b>0.39</b> (2.63) (-4.14]	0.240 (8.75)	<b>0.40</b> (2.37) [-3.56]	
<b>h<sub>0</sub></b>	Speed V weighted by (W <sub>c</sub> +w <sub>c</sub> ) shares $(t=0)$ $[t=1]$	0.036 (0.62)		0.068 (2.18)	<b>2.11</b> (1.25) [0.69]	0.016 (1.44)	<b>3.30</b> (0.47) [0.33]	0.022 (1.16)	<b>2.81</b> (0.53) [0.34]	
<b>t</b>	Time since last regeneration (agerail) $(t=0)$	-0.095 (-0.88)		0.022 (0.26)	<b>0</b>	0.053 (0.63)	<b>0</b>	0.062 (0.80)	<b>0</b>	
<b>R<sub>1</sub></b>	Same region: $\rho_1$ $(t=0)$								<i>0.590</i> <i>(6.91)</i>	
	<b>Log likelihood</b>	<b>-6977.54</b>		<b>-6866.10</b>		<b>-6856.14</b>		<b>-6824.94</b>		
	Number of $\beta_k$ estimated	7		7		9		9		
	Number of $\lambda_k$ estimated	0		0		3		3		
	Difference in degrees of freedom	0		4		6		6		
	Variant run number	95		94		934		937/139		

ii) *the role of control variables and the infrastructure manager's speed of adjustment.* In the third column, the addition of the *Service Target* and *Trajectory Correction* control variables in accordance with (1-B) yields the expected positive and negative signs for these variables —with acceptable  $t$ -statistics— and a significant gain of 10 points in Log Likelihood for a difference of only 2 degrees of freedom.

But the linear<sup>11</sup> forms of these added terms used are the lagged values of target<sup>12</sup> track condition service levels [ $E(S_{ct-1}) - E(S_{ct-2})$ ] and of the difference between observed and target service [ $S_{t-2} - E(S_{ct-2})$ ], both selected primarily because they are more statistically significant<sup>13</sup> than contemporaneous values

<sup>10</sup> We assumed that trains are driven at the lowest of 90% of maximum allowed speed  $v$  on the segment and of the maximum speed allowed by train type, assumed to be 200 km/h for main intercity (GL) and regional (TER) trains, 100 km/h for freight (FRT) trains and 50 km/h for locomotives (HLP).

<sup>11</sup> BCT cannot be used because these variables are not strictly positive.

<sup>12</sup> For the derivation and estimation of target track condition  $Sc$ , see source paper.

<sup>13</sup> The correlation between current and lagged Target Service variables is -0,21 and that between Trajectory Correction variables is 0,81; one could envisage using both current and lagged Trajectory correction variables if Maintenance cost data were available for more than the single year 2007.

$[E(Sc_t) - E(Sc_{t-1})]$  and  $[S_{t-1} - E(Sc_{t-1})]$ . Reasonably, lagged values imply a response that, for various reasons including the discrete yearly nature of the formulation and of data, is not immediate.

Other elasticities are also reasonable and stable except for that of *segment length*, unfortunately much too high and significant until the introductions of spatial correlation in Column 4. In particular, the total ton traffic variable ( $W+w$ ) is highly significant with BCT values between 0,39 and 0,63 resembling the 0,49 power for the axle weight variable in the referenced UK ORR formula for track damages.

iii) *directed correlation of residuals: within region effects*. The addition of first order directed autocorrelation in Column 4 yields huge further gains (31 points of Log Likelihood for the single additional parameter  $\rho_1=0,59$ ) under the assumption that error terms of segments are correlated if they belong to the *same SNCF administrative region*. As indicated in Table 3, the number of track segments present in the sample varies across the 23 administrative regions of 2007, with numerical implications for the structure<sup>14</sup> of  $R_1$ .

**Table 3. SNCF regions: neighbors (nb), segments (sg) in the sample and *t*-statistic of dummies**

N°	Region*	nb	sg	<i>t</i> -stat.***	N°	Region*	nb	sg	<i>t</i> -stat.***
10	PARIS EST**	5	5	reference	43	BORDEAUX	4	51	(0.10)
14	REIMS	7	17	(0.02)	44	LIMOGES	4	17	(0.28)
17	METZ-NANCY	3	36	(0.11)	46	TOURS	6	19	(0.09)
18	STRASBOURG	2	31	(-0.47)	47	TOULOUSE	4	33	(-0.03)
20	PARIS NORD**	7	25	reference	50	PARIS SUD EST**	8	24	reference
23	LILLE	2	38	(0.34)	53	DIJON	7	30	(0.37)
24	AMIENS	4	14	(0.10)	54	LYON	5	35	(0.66)
30	PARIS ST-LAZARE**	5	11	reference	56	CLERMONT-FERRAND	7	27	(0.61)
33	ROUEN	5	21	(0.23)	57	CHAMBERY	3	31	(0.69)
34	PARIS RIVE GAUCHE**	5	25	reference	58	MARSEILLE	3	29	(0.39)
36	RENNES	2	14	(-0.17)	59	MONTPELLIER	4	23	(-3.31)
37	NANTES	5	24	(-0.16)	NUMBER OF SEGMENTS AND OF DUMMIES			580 sg	18 dummies
* In 2012, a merger of regions 10 and 14 and of regions 20 and 24 reduced the total number to 21.									
** The pie-shaped administrative regions located in Île-de-France are treated as first neighbors of one another in matrix R <sub>2</sub> .									
*** In run 140, <i>t</i> -stat. of 18 dummy variables added to Column 4 (run 139), increasing Log Likelihood from -6824.94 to -6817.76.									

iv) *other forms of directed autocorrelation*. A systematic attempt (with a slightly larger data set of 673 observations) to define a second type of directed autocorrelation  $R_2$  based on regions sharing a border with that of the given segment, a number which varies from 2 to 8 as indicated in Table 3, and to define yet another type  $R_3$  under the assumption that the error term of a segment could be correlated with those of other segments sharing one or more train lines, yielded no significant correlation in models specified as those found in Columns 3 and 4 of Table 2. The attempt to further estimate matching proximity<sup>15</sup> parameters  $\pi_2$  and  $\pi_3$  in addition to  $\rho_2$  and  $\rho_3$ , in accordance with Eq. (2-E), also failed to provide statistically significant gains in fit beyond those of models 3 or 4, whether the additional AR-C-D processes were considered separately or jointly with  $\rho_1$ .

In this latter  $R_3$  case, the fact that segment residuals are uncorrelated with use by shared train lines either close (*i.e.*  $\rho_3$  estimates do not differ from 0) or far (*i.e.*  $\pi_3$  estimates do not differ from 1) suggests that the underlying representation of the track damage function in Eq. (1-B) is adequate.

Everything considered, the hypothesized role of within-region SNCF administrative practices appears considerable in explaining reported total maintenance cost by segment; and unaccounted for influences of neighboring regions have a negligible effect. We now wish to probe this stunning improvement in fit obtained with  $R_1$  in Column 4 and try to make some sense of it.

<sup>14</sup> Consider Region 10, with the smallest number of segments, and the Region 43 with the highest: in matrix  $R_1$ , the 5 rows corresponding to the former have 4 values of 1,00 and the 51 rows corresponding to the latter have 50 such values. In the row-normalized  $\tilde{R}_1$ , the former values become 1/4 and the latter 1/50.

<sup>15</sup> A parameter  $\pi_1$  would not be meaningful: the matrix  $R_1$  partitioned by region is symmetric and block diagonal. It is not clear what its powers would mean.

#### 4. Within-region correlation among regression errors

**The near-demise of segment length as an explanatory variable.** Observations on maintenance expenses per segment result from administrative procedures assigning track work invoices to segments. A natural reaction to the existence of within-region correlation among residuals, positive or negative, is then to think that it reflects some systematic arbitrariness in the assignment of these amounts across track segments of any given region, a view partly sustained by noting that taking correlation into account in Column 4 of Table 2 reduces both the size of the coefficient of the *segment length* variable and its statistical significance to a low enough level expected from the start<sup>16</sup>. But even granting this, the positive sign of the  $\rho_1$  estimate still requires interpretation.

**Substitution/complementarity interpretation.** A first candidate interpretation of that positive sign is that segments within the same region behave as substitutes. To see why  $\rho > 0$  could result from the treatment of segments as *substitutes*, consider the following, based on applying Eq. (2-B)-(2-C) at the level of two observations, numbered 2 and 4, for which one writes:

$$(3-1) \quad y_2 = \beta X_2 + v_2 \quad \text{and} \quad y_4 = \beta X_4 + v_4,$$

with

$$(3-2) \quad v_2 = \rho v_4 + w_2,$$

which, after replacement of  $v_2$  in (3-1), of  $v_4$  by  $y_4 - \beta X_4$ , and a rearrangement of terms, leads to:

$$(3-3) \quad y_2 = \rho y_4 + \beta(X_2 - \rho X_4) + w_2.$$

If  $\rho > 0$ , the impact of  $X_4$  on  $y_2$  in this last relationship will be in the opposite direction from that of  $X_2$ , as occurs between substitute goods in problems where  $X_2$  and  $X_4$  denote their prices and the indices pertain not to observations but to quantities demanded, as given by distinct equations. Similarly,  $\rho < 0$  will have the opposite effect, increasing the impact of a unit change in  $X_k$ , as between complements if the  $X_k$  again stand for prices. But this is only a simple behavioral analogy.

**Missing variable interpretation.** An alternate interpretation comes from time-series analysis where positive serial correlation between residuals is often caused by the absence from regressor lists of variables, notably those standing for dynamic effects (Spanos, 1987-1988): if a particular  $X_k$  is missing and its values are correlated over time, then residual  $v_i$  will contain a missing component  $\beta_k X_{k,t}$  that easily generates serial correlation. One way to explore the nature of variables presumably missing here is to add to a reference model regional dummy variables representing locally missing factors and to probe their coefficients (each expressing a difference with the coefficient of the region selected as arbitrary reference) for meaningful identity clues.

On these lines, adding 18 dummy variables to the Column 4 specification (the remaining 5 pie-shaped regions centered on downtown Paris then forming the Île-de-France reference region), increases the Log Likelihood by only 7 points<sup>17</sup>; also, the  $t$ -statistics of the 18 new coefficients, shown in Table 3, obtain 13 positive and 5 negative values, the latter including the single case (59, Montpellier) of a statistically significant difference from the Île-de-France region coefficient,

<sup>16</sup> Unfortunately, national studies produced by the CATRIN consortium (Wheat *et al.*, 2009) all tended to show that segment lengths are statistically significant, a result which may everywhere indicate an endogeneity of segment lengths.

<sup>17</sup> The gain from -6824.94 in run 139 to -6817.76 in run 140 (not shown in Table 2). In previous work on a 1999 database, using a Generalized Box-Cox specification (2-A) applied to (1-A) — with interactions among explanatory variables accounted for (Gaudry & Quinet, 2003) —, we added 22 dummy variables, taking region 10 as the arbitrary reference: this improved the Log Likelihood by 24 points. The comparable exercise carried out here yields hardly any gain, which suggests again that second-wave specification (1-B), estimated on censored observations, is more realistic than first-wave specification (1-A).

possibly because the dryer and warmer climate reduces maintenance costs. Moreover, one notes that these 5 regions with negative dummy signs are geographically dispersed, in contrast with 3 of the regions with positive  $t$ -values around 0,60 (54, Lyon; 56, Clermont-Ferrand; 57, Chambéry), which are contiguous: this timid similarity may perhaps be due to their significant snowfall, absent from the specification, which increases maintenance costs<sup>18</sup>.

**Missing climate variables?** The first clue obtained from this addition of dummy variables is therefore climatic, but it is a weak clue. Firstly because of the small explanatory power of the additional 18 dummy variables and secondly because the autocorrelation is only slightly weakened: the estimated value of  $\rho_1$  decreases from 0,59 to 0,44 but further removing this autocorrelation parameter (in run 138, not shown) massively lowers the Log Likelihood by 10 points: regional dummies then, of little interest in terms of fit, are not a substitute for within-region correlation — still in need of an interpretation. Displacing the mean error does not affect correlation among errors.

The second reason for doubting the importance of climatic variables is that, in a model that did not contain a Cumulative Traffic variable  $W$ , Rail Concept (2011, p. 46, Test 4) found only weak impacts of geographic dummy variables defined by climatic zone<sup>19</sup>: oceanic, semi-oceanic, continental, mountainous, Mediterranean. The question cannot be solved without trying climatic variables within the new model formulation containing  $W$ , but the above dummy variable tests make the perspective unpromising.

**Differences of autocorrelation coefficients across administrative regions.** To test the assumption, maintained up to this point, that the 23 within-region correlation coefficients are all equal, we grouped 21 administrative regions into 4 larger geographical units (East, North, South-West and South-East) detailed in Table 4, each containing a sufficient numbers of segments for estimation purposes and including its Paris station head region.

**Table 4. Four geographic partitions of matrix  $R_1$  into pairs of complementary matrices**

		Geographic partitions								Average segment length (km)		Average number of switches per segment	
Impact matrix criterion:		East		North		South-West		South-East					
N°	Administrative region	East	Other	North	Other	South-West	Other	South-East	Other				
10	PARIS EST	5			5		5		5	20,38	25,59	72,60	32,36
14	REIMS	17			17		17		17	31,93		25,53	
17	METZ-NANCY	36			36		36		36	29,78		32,94	
18	STRASBOURG	31			31		31		31	18,07		28,94	
20	PARIS NORD		25	25			25		25	21,90	22,31	42,20	30,36
23	LILLE		38	38			38		38	18,14		21,53	
24	AMIENS		14	14			14		14	34,35		33,21	
30	PARIS ST-LAZARE		11		11		11		11	7,36		1,64	
33	ROUEN		21		21		21		21	32,43	23,81	28,19	19,06
34	PARIS RIVE GAUCHE		25		25	25			25	25,52	34,51	33,68	25,52
36	RENNES		14		14	14			14	43,95		30,07	
37	NANTES		24		24	24			24	32,37		23,58	
43	BORDEAUX		51		51	51			51	34,83		21,67	
44	LIMOGES		17		17	17			17	39,21		23,47	
46	TOURS		19		19	19			19	33,25		32,79	
47	TOULOUSE		33		33	33			33	36,67		21,64	
50	PARIS SUD EST		24		24		24	24		35,84		32,10	
53	DIJON		30		30		30	30		37,16	30,77		
54	LYON		35		35		35	35		32,65	36,31		
56	CLERMONT-FERRAND		27		27		27	27		28,39	16,22		
57	CHAMBERY		31		31		31	31		29,21	26,29		
58	MARSEILLE		29		29		29	29		23,49	30,62		
59	MONTPELLIER		23		23		23	23		39,90	29,61		
	NUMBER OF SEGMENTS	89	491	77	503	183	397	199	381	28,22			25,48
	MATRIX NAME	RE	RNE	RN	RNN	RRG	RNRG	RSE	RNSE	R <sub>1</sub>	--	--	--

<sup>18</sup> In Sweden, as much as 80% of track maintenance costs are attributable to snow clearing (Andersson, 2006).

<sup>19</sup> Rail Concept also tested the impact of tri-dimensional location (latitude, longitude, altitude), as well as that of frost, precipitations and relative humidity but they were not statistically significant.

These 4 geographical units, jointly accounting for all administrative regions except two (30, Paris St-Lazare; 33, Rouen) with too few observations together to yield robust estimates of own autocorrelation coefficients, are then used to define 4 pairs of complementary sub-sets of  $R_1$ , *i.e.* of sub-matrices  $R_D$  and  $R_R$ , all documented in Table 4.

The results of these tests, shown in Table 5, demonstrate that there in fact exists no correlation among residuals within the East and North groupings: not only does one make significant gains (of 3.6 and 2.2 points, respectively) in Log Likelihood by using two orders of autocorrelation for these groupings, but the estimated small own correlation coefficients are negative<sup>20</sup> but not significantly different from 0. By contrast, this kind of splitting has no impact whatsoever for the South-West and South-East groupings where all correlation coefficients remain statistically indistinguishable from the maintained value of 0.590 found in the reference model, as demonstrated by the trivial gains in Log Likelihood (of less than one point) yielded by these pairs of splits of  $R_1$ . Moreover, the union of East and North matrices yields gains of 6.05 points, *i.e.* a value close to the sum of the gains indicated for each grouping considered separately, and estimates of autocorrelation coefficients compatible with Minkowsky's triangle inequality condition that  $|\rho_1| + |\rho_2| \leq 1$ , implying stationarity.

**Table 5. Testing geographic and sociological decompositions of matrix  $R_1$**

		Table 2 Col. 4	East	North	East & North	South- West	South- East	Family structure
<b>Log Likelihood</b>		<b>-6824.94</b>	<b>-6821.35</b>	<b>-6822.75</b>	<b>-6818.89</b>	<b>-6424.90</b>	<b>-6824.21</b>	<b>-6824.89</b>
$R_D$	$\rho_D$ ( $t=0$ )		-0.536 (-0.26)	-0.320 (-0.38)	-0.117 (-0.66)	0.617 (4.25)	0.676 (5.08)	0.612* (5.53)
$R_R$	$\rho_R$ ( $t=0$ )		0.627 (7.37)	0.617 (7.19)	0.698 (6.25)	0.577 (5.35)	0.525 (4.85)	0.565** 4.45
$R_1 \equiv R_D \cup R_R$	$\rho_1$ ( $t=0$ )	0.590 (6.91)						
<b>Run number</b>		<b>937/139</b>	<b>146</b>	<b>149</b>	<b>249</b>	<b>155</b>	<b>152</b>	<b>143</b>
*Liberal ; **Autoritarian.								

The absence of correlation among residuals of Eastern and Northern groupings of regions considered separately or together, combined with strongly resilient positive correlation within regions of the rest of France, is puzzling. The 4 right-hand side columns of Table 4 indicate that segments in those East and North regions are shorter and contain more switches than the average segment in the sample. But these two variables have already been accounted for in the explanations of the total maintenance cost per km, as made clear in Table 2. Moreover, as implied in Section 2.2 above, it would be unwise to make the elements of  $R_1$  functions of continuous variables, such as segment length or switches per segment, because the resulting estimates of  $\rho_i$  then become conditional on units of measurement of the selected continuous variable (Bolduc *et al.*, 1989, p. 369), an unacceptable dependency.

**Hidden cultural factors?** The SNCF is a centralized firm with «horizontal» accounting controls applied across all administrative regions: a finding of significant differences between the East and North groupings of regions and the rest of France then calls for an explanation. Could residuals within Eastern and Northern groupings be uncorrelated because of differences in accounting practices imputable to local cultural factors? Consider regional firm and family traditions.

<sup>20</sup> Splitting  $R_1$  has no impact on the segment length variable elasticity and on the  $t$ -statistic of its coefficient; in fact, neither does it have any impact on any of the other coefficients. Note that estimates for the North grouping satisfy Minkowsky's triangle inequality condition for stability, but that this is not the case for all other groupings. The implications of this fact in spatial models seems to be even more difficult to ascertain than in time series models. In our examples, estimated elasticities do not seem to differ between the stable and unstable cases.





One possibility is that accounting traditions differed in the original regional railway firms merged into the SNCF conglomerate in 1938. Indeed, the East grouping closely matches the network of the original *Compagnie de l'Est*, including its section later managed by *l'Administration des chemins de fer d'Alsace et de Lorraine*<sup>21</sup>; and the North grouping resembles the network of the *Compagnie du Nord*. The finding concerning the Eastern grouping has nothing to do with the fact that, in Alsace-Lorraine, rail traffic runs on the opposite track from that used in the rest of France! Rather, could it be that, in both East and North cases, firm-specific accounting traditions have long lives?

On this point, consulting firms<sup>22</sup> hired in 2011 by *Réseau Ferré de France* (RFF) and the regional transport authority *Syndicat des transports d'Île-de-France* (STIF) to study the value of train regularity found that administrative forms used to register the causes of late train arrivals varied across the main large stations of Paris, each using its reporting formats defined before 1936. Our historical hypothesis concerning the similar survival of distinct accounting practices could only be confirmed by a detailed study, well beyond the scope of this paper, that might also explain in passing differences in average management section length.

Another testable possibility, more sociological, postulates that local family structures might influence employee attitudes towards centralized administrative norms and affect the assignment of work invoices to segments. To test this wild card assumption, we use Todd's (1990) map of prevailing family structures found in Table 6: it is defined in terms of two categorical dimensions, namely amount of parental authority and inequality in the treatment of siblings.

**Table 6. Family structure in France and the Liberal/Authoritarian partition of SNCF regions**

Les types familiaux

	<i>Rapports entre parents et enfants</i>	<i>Rapports entre frères</i>
	Nucléaire égalitaire	
	Souche	
	Communautaire	
	Nucléaire absolu	
	libéraux	égalitaires
	autoritaires	inégalitaires
	autoritaires	égalitaires
	libéraux	non égalitaires

Partition according to parent-children relationship

N°	Region	Segments	Liberal	Authoritarian
10	PARIS EST	5	5	
14	REIMS	17	17	
17	METZ-NANCY	36		36
18	STRASBOURG	31		31
20	PARIS NORD	25	25	
23	LILLE	38		38
24	AMIENS	14	14	
30	PARIS ST-LAZARE	11	11	
33	ROUEN	21	21	
34	PARIS RIVE GAUCHE	25	25	
36	RENNES	14		14
37	NANTES	24		24
43	BORDEAUX	51		51
44	LIMOGES	17	17	
46	TOURS	19	19	
47	TOULOUSE	33		33
50	PARIS SUD EST	24	24	
53	DIJON	30	30	
54	LYON	35		35
56	CLERMONT-FERRAND	27		27
57	CHAMBERY	31		31
58	MARSEILLE	29		29
59	MONTPELLIER	23		23
	SEGMENTS	580	208	372
	MATRIX DEFINITION	R1	R1.L	R1.A

\*Source: Todd (1990, p. 88).

\*Source: Todd (1990, p. 88).

We obtain a good match between areas of Authoritarian or Liberal parental attitudes and SNCF administrative regions, but find it impossible to achieve a 4-split partition of the latter also matching differences in the degree of hierarchy among siblings. As defined in any case, this sociological hypothesis concerning subtle influences of family structure on employee attitudes turns out to be useless, as the results of the last column of Table 5 make clear.

<sup>21</sup> We tried to split  $R_1$  between Alsace-Lorraine (regions 18 and 17) and the rest of France, but the conditioning of the complementary matrix for the rest of France did not allow for a stationary value of its correlation coefficient.

<sup>22</sup> Stratec, a Belgium firm, and Significance, a Dutch firm.

## 5. Conclusion

Using as reference a recent France-wide model of rail infrastructure maintenance cost where regression model residuals associated to track segments are assumed to be similarly correlated among themselves within the 23 administrative regions of the national firm (SNCF), we attempt to explain the presence of the strong positive and stationary correlation coefficient estimated<sup>23</sup> in this manner and to probe the extent to which the assumption of a common correlation coefficient across administrative regions might be refined and interpreted.

We first find that country-wide within-region correlation among residuals is not weakened if regional dummy variables are added to the model in the hope of finding interpretative clues, notably of the presence of climate effects unaccounted for in the model specification. The implicit region-specific weather effects indirectly so represented turn out to be extremely weak, if present at all, and do not affect the strength of extant within-region correlation or the need to make sense of it.

We then explore differences in correlation coefficient estimates among regions and show that, within the reference maintenance cost model, two large geographic groupings of regions, each comprising in the East or in the North about 15% of total available track segments, in fact have own residuals that are uncorrelated among themselves, in contrast to the more numerous 70% of segment residuals remaining in the rest of regions, which as a group remain robustly and positively correlated, and in a stationary manner (with  $0 \leq \rho_t \leq 1$ ). As the two groupings with uncorrelated residuals closely correspond to the networks of regional firms merged into the SNCF conglomerate in 1938, we hypothesize, *faute de mieux* and in the absence for the moment of refined local climate variables to pursue unpromising missing variable tests, that within-firm accounting traditions might have survived centralized management control of the assignment of track surveillance, maintenance and repair invoices to track sections centrally defined for accounting management purposes.

## 6. References

- Andersson, M. (2006). *Marginal railway infrastructure cost estimates in the presence of unobserved effects*. Annex to Deliverable D3 of GRACE (Generalisation of Research on Accounts and Cost Estimation). ITS, University of Leeds, Leeds, 34 pages.
- Andersson, M. (2009). CATRIN (Cost Allocation of TRansport INfrastructure cost), Deliverable 8, Rail Cost Allocation for Europe – Annex 1A, *Marginal Costs of Railway Infrastructure Wear and Tear for Freight and Passenger Trains in Sweden*. VTI, Stockholm, January, 33 pages.
- Andersson, M., Smith, A., Wikberg, A., Wheat, P. (2012). Estimating the marginal cost of Railways Track Renewals using corner solution models. *Transportation Research A* 46, 6, 954-964.
- Ben-Akiva, M., Gopinath, D. (1995). Modelling infrastructure performance and user costs. *Journal of Infrastructure Systems* 1, 1, 33-43.
- Berndt, E.R., Khaled, M.S. (1979). Parametric Productivity Measurement and Choice among Flexible Functional Forms. *Journal of Political Economy* 87, 6, 1220-1245, December.
- Blum, U.C.H., Bolduc, D., Gaudry, M.J.I. (1995/1996). From Correlation to Distributed Contiguities: a Family of AR-C-D Autocorrelation Processes. *Zeitschrift für Regionalwissenschaft* 16/17, 99-128.
- Bolduc, D. (1987). *On the estimation of models with generalized SAR(1) processes on the residuals of a regression*. Publication CRT-508, Centre de recherche sur les transports, Université de Montréal.

---

<sup>23</sup> This estimated coefficient was accepted, but other possible sources of correlation were rejected, notably the presence of neighbouring influences across region boundaries and that of shared train lines influencing infrastructure costs in physical ways unaccounted for in the reference model.



- Bolduc, D., Dagenais, M.G., Gaudry, M.J.I. (1989). Spatially Autocorrelated Errors in Origin-Destination Models: A New Specification Applied to Aggregate Mode Choice. *Transportation Research B* 23, 5, 361-372.
- Borger, B.de (1991). Hedonic versus homogeneous output specifications of railroad technology: Belgian railroads 1950-1986, *Transportation Research* 25A, June, 227-238.
- Borger, B.de (1992). Estimating a Multiple-Output Generalized Box-Cox Cost Function: Cost Structure and Productivity Growth in Belgian Railroad Operations, 1950-86. *European Economic Review* 36, pp. 1379-1398.
- Caves, D.W., Christensen, L.R., Tretheway, M.W. (1980a). Flexible Cost Functions for Multiproduct Firms. *Review of Economics and Statistics*, 477-481 (Notes Section).
- Caves, D.W., Christensen, L.R., Swanson, J.A. (1980b). Productivity in U.S. Railroads, 1951-1974. *Bell Journal of Economics* 11, 1, 168-181.
- Caves, D.W., Christensen, L.R., Tretheway, M.W., Windle, J. (1985). Network effects and the measurement of returns to scale and density for U.S. railroads. Ch. 4 in A.F. Daughety (ed.), *Analytical Studies in Transport Economics*, 97-120, Cambridge University Press, Cambridge.
- Cliff, A.D., Ord, J.K. (1981). *Spatial Processes: Models and Applications*. Pion Limited, London.
- Christensen, L.R., Jorgenson, D.W., Lau, L.J. (1971a). Conjugate Duality and the Transcendental Logarithmic Function. *Econometrica* 39, 4, 255-256, July.
- Christensen, L.R., Jorgenson, D.W., Lau, L.J. (1971b). Transcendental Logarithmic Production Frontiers. *Review of Economics and Statistics* 55, 1, 28-45.
- Gaudry, M. (2004). The Robust Gravity Form in Transport and Trade Models. *Zeitschrift für Kanada-Studien* 24, 1, Band 44, 144-161.
- Gaudry, M.J.I., Blum, U.C.H. (1988). An Example of Correlation Among Residuals in Directly Ordered Data. *Economics Letters* 26, 4, 335-340.
- Gaudry, M., Laferrière, R. (1989). The Box-Cox Transformation: Power Invariance and a New Interpretation. *Economics Letters* 30, 27-29.
- Gaudry, M., Quinet, É. (2003). *Rail track wear-and-tear costs by traffic class in France*. Publication AJD-66, Agora Jules Dupuit, Université de Montréal, [www.e-ajd.net](http://www.e-ajd.net), et W.P. 04-02, Centre d'Enseignement et de Recherche en Analyse Socio-économique, École nationale des Ponts et Chaussées (ENPC-CERAS), [www.enpc.fr/ceras/labo/accueil.html](http://www.enpc.fr/ceras/labo/accueil.html), 26 pages, November. Paper presented at the First Conference on Railroad Industry Structure, Competition and Investment, Institut D'Économie Industrielle (IDEI), Toulouse.
- Gaudry, M., Quinet, É. (2009a). CATRIN (Cost Allocation of TRansport INfrastructure cost), Deliverable 8, Rail Cost Allocation for Europe – Annex 1Di– *Track Maintenance Costs in France*. Funded by Sixth Framework Programme. VTI, Stockholm.
- Gaudry, M., Quinet, É. (2009b). *Track Wear and Tear Cost by Traffic Class: Functional Form, Zero-Output Levels and Marginal Cost Pricing Recovery on the French Rail Network*. Publication AJD-130, Agora Jules Dupuit, Université de Montréal et Working Paper N° 2009-32 Paris-Jourdan Sciences Économiques, 46 pages, Septembre. Forthcoming as Ch. 7 in Gaudry, M., Tegnér, G. (eds), *Transport Demand and Cost Models: The role of Box-Cox transformations*. Elsevier Science, Oxford.
- Gaudry, M., Blum, U., Liem, T. (2000). Turning Box-Cox, including Quadratic Forms in Regression. In Gaudry, M., Lassarre, S., (eds), *Structural Road Accident Models: The International DRAG Family*, Pergamon, Elsevier Science, Oxford, Ch. 14, 335-346.
- Gaudry, M., Lapeyre, B., Quinet, É. (2013). *Infrastructure maintenance, regeneration and service quality economics: A rail example*. Publication AJD-136-E, Agora Jules Dupuit, Université de Montréal, [www.e-ajd.net](http://www.e-ajd.net); Working Paper N° 2011-03, Paris-Jourdan Sciences Économiques, 42 p. juillet 2013.
- Idström, T. (2002). *Reforming the Finnish railway infrastructure charge – an econometric analysis of the marginal costs of rail infrastructure use*. In Finnish. Masters Thesis. Department of Economics. University of Jyväskylä.

- Johansson, P., Nilsson, J.-E. (2002). *An Economic Analysis of Track Maintenance Costs*. Deliverable 10 Annex A3 of UNITE (UNification of accounts and marginal costs for Transport Efficiency), Funded by EU 5th Framework RTD Programme. ITS, University of Leeds, Leeds, 23 p.. Online: <http://www.its.leeds.ac.uk/projects/unite/>.
- Johansson, P., Nilsson, J.-E. (2004). An economic analysis of track maintenance costs. *Transport Policy* 11, 277-286.
- Koyck, L. (1954). *Distributed Lags and Investment Analysis*. North Holland, Amsterdam.
- Link, H. (2009). CATRIN (Cost Allocation of TRAnsport INfrastructure cost), Deliverable 8, Rail Cost Allocation for Europe – Annex 1C - *Marginal costs of rail maintenance in Austria*. VTI, Stockholm, February, 23 pages.
- Marti, M., Neuenschwander, R. (2006). *Case study 1.2E: Track Maintenance Costs in Switzerland*. Annex to Deliverable 3 of GRACE (Generalisation of Research on Accounts and Cost Estimation). University of Leeds, 23 pages.
- Marti, M., Neuenschwander, R., Walker, P. (2009). CATRIN (Cost Allocation of TRAnsport INfrastructure cost) Deliverable 8, Rail Cost Allocation for Europe – Annex 1B. *Track maintenance and renewal costs in Switzerland*. VTI, Stockholm, February, 38 pages.
- Ongaro, D., Iwnicki, S. (2009). CATRIN (Cost Allocation of TRAnsport INfrastructure cost), D8 – Rail Cost Allocation for Europe - Annex 2 - *Railway Infrastructure Wear and Tear for Freight and Passenger Trains in Sweden*. VTI, Stockholm, 21 pages, February.
- Ord, K. (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association* 70, 349, 120-126.
- Rail Concept (2011). *Modèle de coût des infrastructures, Phase 1, Livrable No 6. Présentation des modèles économétriques des coûts de maintenance (entretien/regeneration) et d'exploitation*, Réseau Ferré de France, 77 p., 25 mai.
- Spanos, A. (1987-1988). Error Autocorrelation Revisited: The AR(1) Case. *Econometric Reviews* 6, 2, 285-294.
- Todd, E. (1990). *La nouvelle France. Nouvelle édition mise à jour*. 288 p., Collection Points, Série Politique, No 136, Seuil.
- Tran, C.-L., Gaudry, M. (2008). *LEVEL: The L-2.1 estimation procedures for BC-DAUHESEQ (Box-Cox Directed AUtoregressive HEteroskedastic Single EQuation) regression*. Publication AJD-108, Agora Jules Dupuit, Université de Montréal, 26 p., September 2008. [www.e-ajd.net](http://www.e-ajd.net).
- Wheat, P., Smith, A., Nash, C. (2009). CATRIN (Cost Allocation of TRAnsport INfrastructure cost), Deliverable 8 – *Rail Cost Allocation for Europe*. Funded by Sixth Framework Programme. VTI, Stockholm. 70 pages, March.